Its being used R for at least five years and always been curious about its usage in Brazil. It has been seen some minor personal evidence that the number of users is increasing over time. Conference are also booming. Every year there are at least two or three R conferences in Brazil.

What learned from experience is that software choice is a group decision. It is very likely that you will use whatever your peer group uses. For example, if you are a PhD student, you will never convince your adviser to change research software, even if you have perfectly good reasons!

It takes some independence and autonomy to be able to break free from bad group choices. In academia, you can only do that later on, when you finish your PhD and start your career. Then you can use whatever rocks your boat. And, even for that, it takes courage and humbleness to relearn all you research tricks, from data acquisition to research report.

In this post , the investigation the use of R in Brazil. Rstudio publishes a [log page](http://cran-logs.rstudio.com/) covering all R downloads and package installations, stored in downloadable .csv files. The data is organized by day and very easy to download and parse within R. After downloading it, I organized it by filtering only downloads in Brazil, and saved it in a .rds file. Let’s explore it.

library(tidyverse)

df.dls <- read\_rds('data/r-downloads-brazil.rds')

glimpse(df.dls)

## Observations: 72,853

## Variables: 7

## $ date 2012-10-31, 2012-10-31, 2012-10-31, 2012-10-31, 2012-10…

## $ time 16:17:34, 18:21:35, 19:26:20, 19:28:03, 19:26:03, 19:06…

## $ size 49351035, 33301364, 49351035, 49351024, 1424794, 6652340…

## $ version "2.15.2", "2.15.2", "2.15.2", "2.15.2", "2.15.2", "2.15.…

## $ os "win", "win", "win", "win", "win", "osx", "win", "win", …

## $ country "BR", "BR", "BR", "BR", "BR", "BR", "BR", "BR", "BR", "B…

## $ ip\_id 30, 59, 73, 30, 87, 90, 143, 213, 231, 260, 260, 134, 18…

As you can see, we have the date, time, version, os (platform), country and ip (randomized daily). First of all, let’s see how many downloads per day we have for Brazil. I’m also including the different release dates for major R versions.

df\_by\_day <- df.dls %>%

group\_by(ref.date = date) %>%

summarise(n = n())

df.R.releases <- tibble(ref.date = as.Date(c('2013-04-03', '2014-04-10','2015-04-16',

'2016-05-03', '2017-04-21',

'2018-04-23', '2019-04-26')),

R\_version = c('3.0.0', '3.1.0','3.2.0',

'3.3.0','3.4.0', '3.5.0',

'3.6.0') )

p <- ggplot(data = df\_by\_day, aes(y = n,

x = ref.date) ) +

geom\_point() + geom\_smooth(size = 2) +

labs(x = 'Date (day)', y= 'Number of Downloads',

title = paste0('Number of R downloads in Brazil'),

subtitle = 'Data from Rstudio logs ') +

geom\_vline(data = df.R.releases,

aes(xintercept = ref.date, color = R\_version ), size = 1) +

scale\_color\_grey(start = 0.8, end = 0.2 )

print(p)

The number of downloads is steadily increasing over time. The new releases of R also seems to explain the outliers in the dataset. Let’s clean it a bit by decreasing the frequency and calculating the number of downloads per month, instead of by day.

df\_by\_month <- df.dls %>%

group\_by(ref.month = lubridate::ymd(format(date, '%Y-%m-01'))) %>%

summarise(n = n())

p <- ggplot(data = df\_by\_month, aes(y = n,

x = ref.month) ) +

geom\_point() + geom\_smooth(size = 2) +

labs(x = 'Date (month)', y= 'Number of Downloads',

title = paste0('Number of R downloads in Brazil'),

subtitle = 'Data from Rstudio logs ') +

geom\_vline(data = df.R.releases,

aes(xintercept = ref.date, color = R\_version ), size = 1) +

scale\_color\_grey(start = 0.8, end = 0.2 )

print(p)

Much better! Overall, R downloads average about 910.7 per month, with a monthly compound rate of 6.40%. It means that, each month, the number of downloads is increasing by 6.40% from previous month.

The data also includes information about the operating system. Let’s check its distribution:

df\_by\_os <- df.dls %>%

group\_by(os) %>%

count() %>%

na.omit() %>% ungroup() %>%

mutate(os = fct\_recode(os,

"Windows" = "win",

'Mac OS' = 'osx',

'Linux' = 'src'))

p <- ggplot(df\_by\_os, aes(x = os, y = n)) +

geom\_col() +

labs(x = 'Operation System', y = 'Number of Download Cases',

title = 'Distribution of OS',

subtitle = 'Data from Rstudio logs ')

print(p)

Not unexpectedly, Windows is the winner! I’m very surprised to see that Mac OS presents more downloads than Linux. With an unfavorable exchange rate and many import taxes, the price of a Mac computer (desktop or laptop) are exorbitantly expensive in Brazil. This tells a lot about the purchase power of R users.

I hope you liked this post. Next time I’ll analyze the logs of package installation in Brazil.